Anonymization of data for open science in psychology

Jiří Novák[1,2,3]

[1] University of Zürich

[2] University of Applied Sciences and Arts Northwestern Switzerland

[3] Swiss Data Anonymization Competence Center

1. Background

There is a growing demand for more research data to be made openly available. The reproducibility of findings is in crisis [1], and more openly available data would make research more transparent and accessible.

However, **psychological datasets often include sensitive personal information that necessitates privacy protection**.

OPEN SCIENCE, OPEN ACCESS, OPEN DATA

open science OPEN ACCESS OPEN DATA

Data that results from publicly funded research should be:

- **Findable, Accessible, Interoperable, Reusable** ('**FAIR** principles') [2] [3] therefore replicable, transparent, shareable, trustworthy, verifiable and accountable.
- **As open as possible, as closed as necessary.**

2. Methodology

Released data can provide attackers with new information about specific respondents. For safe dissemination, researchers may use **Statistical Disclosure Control** (SDC) methods [4]:

➤ The traditional approach to protecting data

- **Non-perturbation methods** - partially suppressing or reducing details, e.g. Local suppression, Global recoding, Top and bottom coding, Sampling
- **Perturbation methods** - modifying data, e.g. Adding noise, Record swapping, Microaggregation

➤ **Synthetic data generation** to create artificial data that mimics the original data and can be safely disseminated

- **Joint modeling** - captures entire data distribution simultaneously, e.g. neural networks (GAN)
- **Conditional/sequential modeling** - generates data variable by variable, e.g. parametric (regression) or non-parametric (CART) methods

3. Tutorial: Data Anonymization for Open Science

This introductory tutorial was held at useR! 2024 in Salzburg.

- Statistical disclosure control methods with different anonymization approaches that can be used to protect data confidentiality were shown, and basic concepts of SDC were explained in better detail.
- The usage of packages sdcMicro [5], synthpop [6] and simPop [7] was demonstrated.

Materials to download:

useR! Salzburg 2024

4. Example of synthetic data generation

Let's suppose that we are obliged to share data while reducing the risk that an attacker learns something new about respondents.

➤ **Dataset Description**

- The data for this example is from the Answers to the Machivallianism Test, a version of the MACH-IV from Christie and Geis [8], which comprises 73,489 records.
- Includes variables about Likert-rated items and demographic/other items.

➤ **Anonymization tool**

- Synthetization was performed using the R package **synthpop** [6] with selected method CART.

Data utility

The *utility of synthetic data* is measured by how the results from analyses of synthetic data differ from those derived from the real data [9]. There is a **risk-utility trade-off** in anonymizing data.

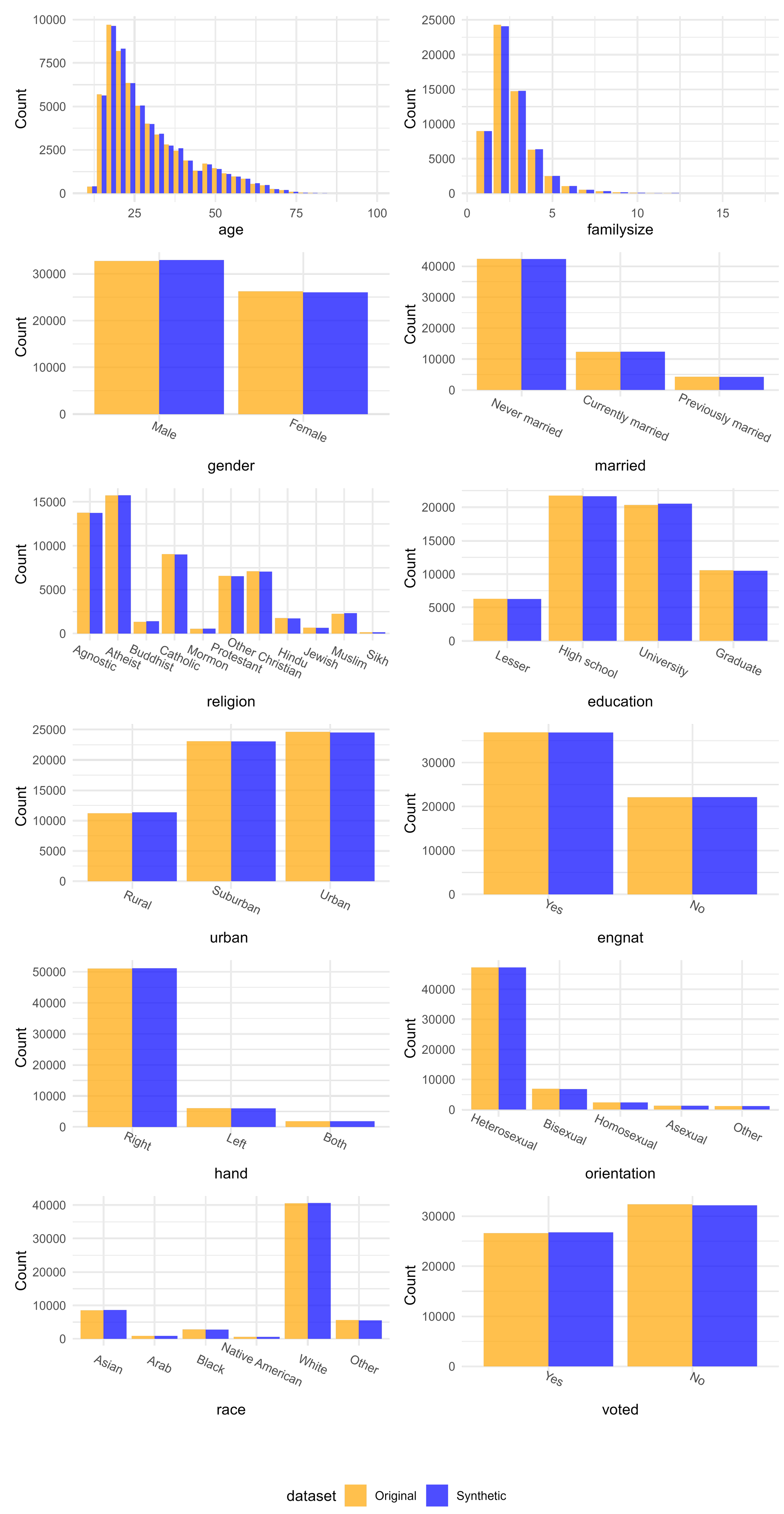

Figure 1: Difference in distribution between Original and Synthetic dataset

The plots compare the marginal distribution in selected variables for both the original and synthetic datasets. The similarity in the histograms and bar plots suggests that the synthetic data maintains the original data's univariate structure well.

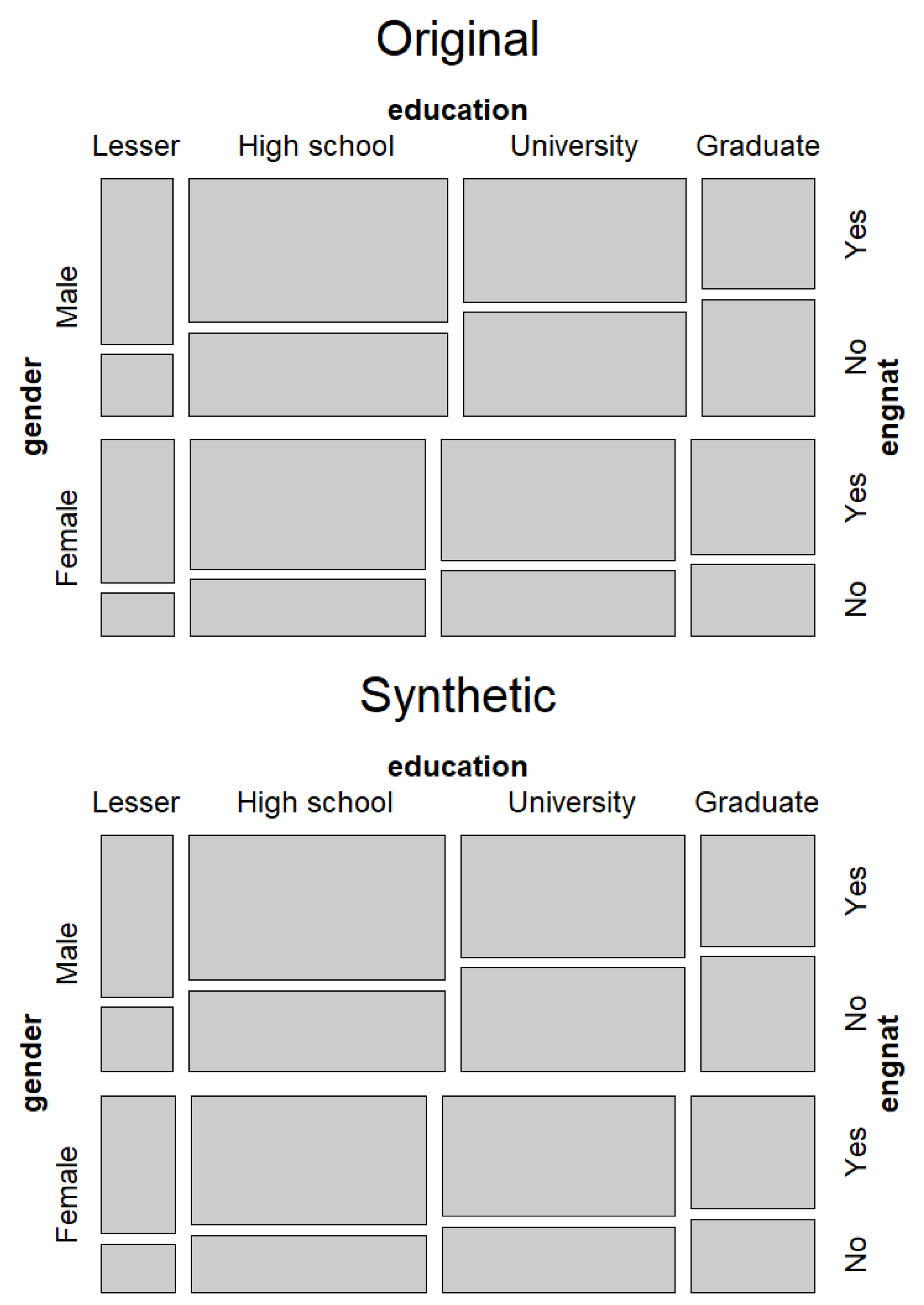

Figure 2: Mosaic plots for selected variables

The mosaic plots display differences in structure for categorical data. In this case, the synthetic and original datasets show highly similar distributions across the variables *gender*, *education*, and *engnat*. This similarity indicates that the synthetic data effectively preserves the relationships and proportions.

5. Forthcoming Research

The goal of our SNSF*-funded project is developing and implementing innovative tools for generating synthetic longitudinal data with a focus on disclosure risk.

References

[1] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251), 2015.

[2] European University Association. The European University Association Open Science Agenda 2025, 2022.

[3] European Commission. Commission recommendation (EU) 2018/790 of 25 april 2018 on access to and preservation of scientific information, 2018.

[4] Matthias Templ. *Statistical disclosure control for microdata*. Springer Berlin Heidelberg, 2017.

[5] Matthias Templ, Alexander Kowarik, and Bernhard Meindl. Statistical disclosure control for micro-data using the R package sdcMicro. *Journal of Statistical Software*, 67(4):1–36, 2015.

[6] Beata Nowok, Gillian M. Raab, and Chris Dibben. synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, 74(11):1–26, 2016.

[7] Matthias Templ, Bernhard Meindl, Alexander Kowarik, and Olivier Dupriez. Simulation of synthetic complex data: The R package simPop. *Journal of Statistical Software*, 79(10):1–38, 2017.

[8] Richard Christie and Florence L. Geis. Answers to the machivallianism test, a version of the MACH-IV, https://openpsychometrics.org/_rawdata/, 2019.

[9] Joshua Snoke, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(3), 2018.

*Acknowledgments

This work was funded by the Swiss National Science Foundation (SNSF) with grant number 211751: *"Harnessing event and longitudinal data in industry and health sector through privacy preserving technologies"*.